

---

# Learning Multiplicative Interactions with Bayesian Neural Networks for Visual-Inertial Odometry

---

Kashmira Shinde<sup>1</sup> Jongseok Lee<sup>1</sup> Matthias Humt<sup>1</sup> Aydin Sezgin<sup>2</sup> Rudolph Triebel<sup>1</sup>

## Abstract

This paper presents an end-to-end multi-modal learning approach for monocular Visual-Inertial Odometry (VIO), which is specifically designed to exploit sensor complementarity in the light of sensor degradation scenarios. The proposed network makes use of a multi-head self-attention mechanism that learns multiplicative interactions between multiple streams of information. Another design feature of our approach is the incorporation of the model uncertainty using scalable Laplace Approximation. We evaluate the performance of the proposed approach by comparing it against the end-to-end state-of-the-art methods on the KITTI dataset and show that it achieves superior performance. Importantly, our work thereby provides an empirical evidence that learning multiplicative interactions can result in a powerful inductive bias for increased robustness to sensor failures.

## 1. Introduction

For safety critical applications such as autonomous driving, it is crucial for learning approaches to effectively rely on multiple sources of information. In case of a localization problem as an example, learning from complementary sensor types such as cameras and IMUs can bring improved robustness and redundancy when there exists sensor corruptions and failures. Existing methods for learning based VIO rely on different strategies to fuse visual and inertial data, which include so-called *FC-fusion networks* (Clark et al., 2017b; Han et al., 2019) and *soft/hard feature selection networks* (Chen et al., 2019; Almalioglu et al., 2019). Broadly speaking, these strategies typically learn additive interac-

tions for integrating multiple streams of information, i.e., given the features  $\mathbf{u}$  and  $\mathbf{v}$  from different modalities, the additive interactions are modeled as  $f(\mathbf{u}, \mathbf{v}) = \mathbf{W} \text{concat}(\mathbf{u}, \mathbf{v})$  where  $\text{concat}(\mathbf{u}, \mathbf{v})$  is the concatenation of the two features, and  $\mathbf{W}$  are the learned fusion network parameters.

Instead, we present an end-to-end trainable deep neural network for monocular VIO with the focus on how to better exploit sensor complementarity. Specifically, the designed network architecture (i) learns multiplicative interactions (Jayakumar et al., 2020) for the fusion network, which is based on multi-head self-attention layers (section 2.2), and (ii) treat the network weights as probability distribution using scalable *Laplace Approximation* (LA) (MacKay, 1992), resulting in *Bayesian Neural Networks* (Section 2.5). In our experiments, we compete with the current methods that learn additive interactions against the proposed network (section 3.1) and further evaluate the obtained uncertainty (section 3.2). As a result, we demonstrate improved performance in terms of accuracy and robustness. We believe that these results imply that a more elaborated fusion might bring performance improvements for the end-to-end VIO.

## 2. Method

We now describe the proposed architecture in figure 1.

### 2.1. Visual-Inertial Feature Extractors

Two sequential images, i.e.,  $\mathbf{u}_{v1}$ ,  $\mathbf{u}_{v2}$  are stacked together and fed to the optical flow feature extractor (FlowNet Dosovitskiy et al. (2015) or even Sun et al. (2018))  $g_{\text{vision}}$ . It learns geometric feature representation to be able to generalise well in unfamiliar environments as opposed to learning appearance representation. The visual features  $\mathbf{b}_v$  obtained from the final layer of FlowNet can be expressed as:

$$\mathbf{b}_v = g_{\text{vision}}(\mathbf{u}_{v1}, \mathbf{u}_{v2}). \quad (1)$$

Raw inertial measurements  $\mathbf{u}_i$  falling between two consequent image frames, i.e.,  $x, y, z$  components of linear and angular acceleration taken together forming a 6 dimensional vector are passed to an encoder  $g_{\text{inertial}}$ . It is composed of a bidirectional LSTM (Long Short-Term Memory) as done by (Chen et al., 2018) and operates at a rate at which IMU

---

<sup>1</sup>German Aerospace Center (DLR), Institute of Robotics and Mechatronics, Wessling, Germany <sup>2</sup>Institute of Digital Communication Systems, Ruhr-Universität Bochum (RUB), Germany. Correspondence to: Kashmira Shinde <kashmira.shinde23@gmail.com>.

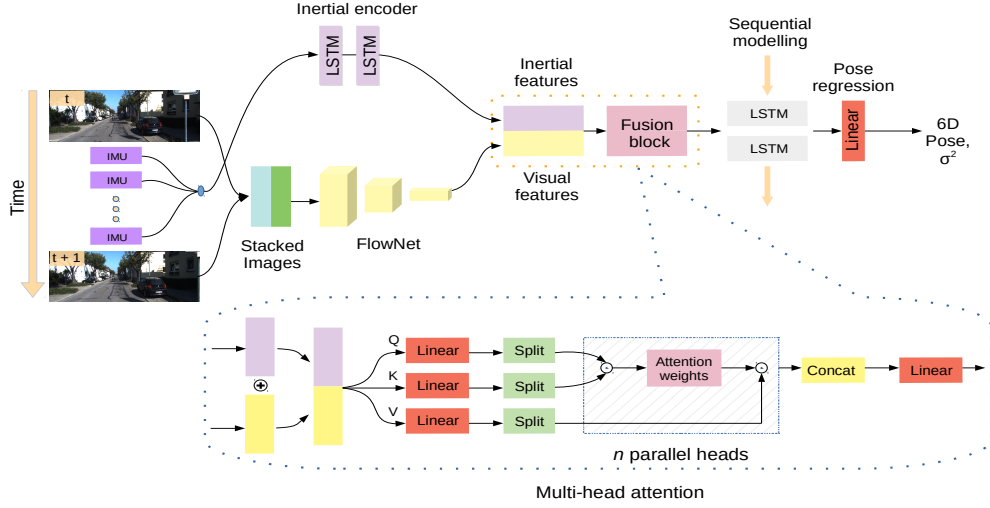


Figure 1. Architecture of the proposed approach.  $\sigma^2$  denotes the variance, i.e., uncertainty of the poses.

receives data. The inertial features  $\mathbf{b}_i$  can be expressed as:

$$\mathbf{b}_i = g_{\text{inertial}}(\mathbf{u}_i). \quad (2)$$

## 2.2. Feature Fusion Mechanism

The fusion block combines both the inertial and visual feature streams generated from the respective feature extractors. Here, a widely used fusion method is to either simply concatenate the two feature streams into a single feature space (Clark et al., 2017b) or re-weight them (Chen et al., 2019), which are so-called additive interactions. In other words, given two different input modalities  $\mathbf{u}$  and  $\mathbf{v}$ , the used fusion blocks rely on rather simple functions: ( $\text{concat}(\mathbf{u}, \mathbf{v})$ ) or a linear layer  $f(\mathbf{u}, \mathbf{v}) = \mathbf{W}_{\text{concat}}(\mathbf{u}, \mathbf{v}) + \mathbf{a}$ .

In contrast to this, we consider so-called *multiplicative interactions* (MI). For deep neural networks, MI can be expressed by the following function (Jayakumar et al., 2020):

$$f(\mathbf{u}, \mathbf{v}) = \mathbf{v}^\top \mathcal{W} \mathbf{u} + \mathbf{v}^\top \mathbf{W}_v + \mathbf{W}_u \mathbf{u} + \mathbf{a}, \quad (3)$$

wherein the weight matrices  $\mathbf{W}_u$  and  $\mathbf{W}_v$ , vector  $\mathbf{a}$  and 3D weight tensor  $\mathcal{W}$  are the learned parameters. A major difference to additive interaction is the product term  $\mathbf{v}^\top \mathcal{W} \mathbf{u}$ . Here, one might argue that neural networks are universal approximators, and thus, adding more costly module may not affect the approximation power of neural networks. Yet, the hypothesis space or the sets of learnable functions, the compactness of the estimator as well as learnability can increase as proved by Jayakumar et al. (2020). When learned correctly, MI can help in preferential selection of hypotheses, and thus may result in an improved inductive bias.

Motivated by this recent result, we proposed to use a more elaborated fusion strategy that can learn MI. To this end, we

propose a fusion block based on multi-head self-attention (denoted by  $f_{\text{attn}}$ ) (Vaswani et al., 2017) which relates different parts of the same sequence to learn a suitable combined feature representation. An advantage over other MI modules is that training can be often easier (e.g. alternatives such as gating networks are often notoriously difficult to train (Shazeer et al., 2017)). In this module, the concatenated visual and inertial features of dimension  $d_k$  are passed to the attention function as query  $\mathbf{q}$  and a key-value pair  $(\mathbf{k}, \mathbf{v})$ . In order to learn distinct representations of the input,  $n$  such attention *heads* perform attention function  $\mathcal{A}$  in parallel for a model of dimension  $N$ . The output from each head  $\mathbf{h}_i$  is then concatenated to produce final attention outputs for a given concatenated feature representation:

$$f_{\text{attn}}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathbf{h} \mathbf{W}^h, \quad (4)$$

where each head  $\mathbf{h}_i$  for  $i = 1, \dots, n$  is given by

$$\begin{aligned} \mathbf{h}_i &= \mathcal{A}(\mathbf{q} \mathbf{W}_i^q, \mathbf{k} \mathbf{W}_i^k, \mathbf{v} \mathbf{W}_i^v) \\ \mathcal{A}(\mathbf{q}, \mathbf{k}, \mathbf{v}) &= \text{softmax} \left( \frac{\mathbf{q} \mathbf{k}^\top}{\sqrt{d_k}} \right) \mathbf{v}, \end{aligned}$$

where  $\mathbf{h} = \text{concat}(\mathbf{h}_1, \dots, \mathbf{h}_n)$  and learned attention weights  $\mathbf{W}_i^q \in \mathbb{R}^{N \times d_k}$ ,  $\mathbf{W}_i^k \in \mathbb{R}^{N \times d_k}$ ,  $\mathbf{W}_i^v \in \mathbb{R}^{N \times d_k}$  and  $\mathbf{W}^h \in \mathbb{R}^{nd_k \times N}$ . The attention outputs  $\mathbf{y}$  obtained from the visual  $\mathbf{b}_v$  and inertial  $\mathbf{b}_i$  feature vectors which are passed to pose regression module can be expressed as:

$$\mathbf{y} = f_{\text{attn}}(\mathbf{b}_v, \mathbf{b}_i). \quad (5)$$

We note that, while the used module is more costly than other simpler variants, we did not find training more difficult.

Furthermore, we use the attention mechanism only in the fusion block, and so, the added number of parameters has smaller contribution to the total number of parameters.

### 2.3. Sequential Modeling and Pose Regression

Next, we describe modeling sequential dependence for ego-motion estimation. In order to learn complex model dynamics (motion model) and to derive connections between sequential features, the core LSTM is used. The resultant representation from the feature fusion block  $\mathbf{y}_t$  at time  $t$  is fed to the core LSTM along with its hidden states from the previous time step  $\mathbf{h}_{t-1}$ . The fully connected layer does the pose regression and the output is a 6D camera pose  $\mathbf{x}_t$ :

$$\mathbf{x}_t = f_{\text{lstm}}(\mathbf{y}_t, \mathbf{h}_{t-1}). \quad (6)$$

### 2.4. Loss Function

The loss as in Wang et al. (2017) constitutes of the summed up mean squared error (MSE) between predicted poses  $\hat{\mathbf{x}}_t = (\hat{\mathbf{z}}_t, \hat{\psi}_t)$  and ground truth  $\mathbf{x}_t = (\mathbf{z}_t, \psi_t)$  (Geiger et al., 2012) where  $\mathbf{z}_t$  denotes position and  $\psi_t$  the orientation at time  $t$ ,

$$\ell = \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^k \left\| \hat{\mathbf{z}}_t - \mathbf{z}_t \right\|_2^2 + \beta \left\| \hat{\psi}_t - \psi_t \right\|_2^2, \quad (7)$$

where  $\beta$  is a scale factor to balance the position and orientation elements for  $M$  mini-batch samples. This scaling factor is to balance the magnitude so that the both translation and rotation are given the similar level of importance.

### 2.5. Uncertainty estimation

We estimate the predictive uncertainty of our method using a LA similar to Ritter et al. (2018); Lee et al. (2020b), but only make use of the diagonal approximation to the Fisher information matrix. Just like MC-dropout (Gal & Ghahramani, 2016) it is a practical method for uncertainty estimation in deep learning without the need to change or retrain the model, but without the reliance on dropout layers.

LA employs information from the inverse curvature (Hessian) of the loss for the covariance matrix of a multivariate Gaussian as an approximation to the model’s weight posterior distribution. We collect the output of multiple stochastic forward passes obtained through sampling from this posterior and perform Bayesian model averaging using Monte Carlo integration to make predictions. Note that we do not consider aleatoric uncertainty as we focus on the influence of model uncertainty on deep sensor fusion. More details can be found in the supplementary material.

## 3. Experimental Results

In this section, we evaluate our approach in comparison to the baselines - DeepVO (Wang et al., 2017), VINet (Clark et al., 2017b) and VIO Soft (Chen et al., 2019). Implementation details can be found in supplementary material.

### 3.1. Analysis of Robustness to Sensor Degradation

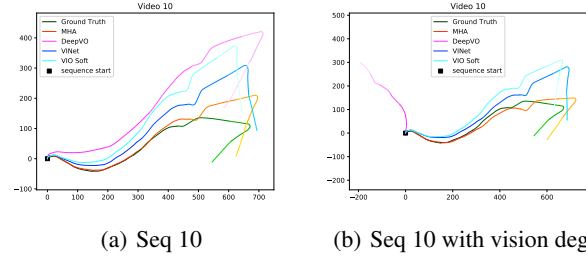


Figure 2. Predicted trajectories of the KITTI dataset comparing MHA and baselines with ground truth (in green). (a) for non-degraded data. (b) for degraded vision data (occlusion, noise+blur, missing images) for sequence 10. Best viewed in colour.

We first compare the baselines with our approach (referred to as MHA) for nominal case where dataset has no degradation. In many real world scenarios, it is plausible for sensors to be imperfect thereby resulting in corrupted data. For a camera, these corruptions can be in the form of occlusions, noisy and blurry images as well as missing image frames. The IMU can have bias in gyroscope data as well as white noise in accelerometer data. To determine the behaviour of the approaches in such scenarios and evaluate the influence of the modalities, we prepare three datasets - a dataset with degraded inertial data, a vision degraded dataset (consisting of image occlusions, blur and missing data) and all sensor dataset (both vision and inertial data are degraded).

Table 1 shows mean values of the sequences while its expanded version can be found in supplementary material. We use the metric provided by Shamwell et al. (2018) - average translation RMSE  $t_{rel}(\%)$  and average rotational RMSE  $r_{rel}(^{\circ}/100m)$  computed on sub-sequences of length 100m - 800m. In our experiments, we find that our method makes predictions with lower RMSE on average, when compared the considered baselines across the sequences. For inertial degradation, it can be seen that as the sensor fusion strategy becomes elaborate for each approach, the robustness to IMU degradation increases. For low to moderate intensity of degradation, the performance of all methods remains similar to nominal case, suggesting visual features are more dominant than inertial ones. For vision degradation case as the dataset is severely degraded, from Figure 2 we see that the performance of DeepVO degrades immensely and tracking is lost completely. Interestingly, when vision degrades,

Table 1. Comparison of approaches for various sensor corruptions are reported along with nominal case. For complexity comparison, the total number of parameters in the networks and fusion blocks are shown. Per sequence results can be found in supplementary material.

MODEL	#TOTAL PARAM	#FUSION PARAM	NOMINAL		INERTIAL		VISION		ALL	
			$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$
DEEPVO	149,518,326	N/A	18.07	5.88	N/A	N/A	FAILS	FAILS	N/A	N/A
VINET	294,828,726	<b>31,020</b>	13.98	4.66	16.31	5.28	15.46	4.97	15.87	5.14
VIO SOFT	<b>58,969,014</b>	262,656	13.65	4.83	14.05	4.94	14.59	4.58	15.48	4.63
MHA	60,019,638	1,313,280	<b>11.42</b>	<b>3.94</b>	<b>12.48</b>	<b>3.93</b>	<b>12.61</b>	<b>3.98</b>	<b>12.47</b>	<b>3.79</b>

$t_{rel}(\%)$ : translational RMSE,  $r_{rel}(^\circ/100m)$ : rotational RMSE.

the translation error increases for all the approaches, but the rotational error remains comparable to the nominal case. This suggests that the visual features might contribute to determining translation and inertial contribute to rotation, and furthermore, inertial features might become more reliable in case of strong visual degradation.

To summarize, exploiting the combination of sensors can be beneficial for odometry in comparison to use of a single sensor (DeepVO). Among all three sensor fusion approaches, our approach being more expressive than VINet and VIO Soft, performs better than them on average in terms of accuracy (lower translational and orientation error) in our experiments. Our method appears to be more robust to sensor corruptions and tends to diverge less as compared to the other two approaches. We further report total number of parameters in a network, and also in the fusion block. As expected, our approach shows an increased cost for the fusion parameters, but requires not as much total network parameters to obtain an improved performance, suggesting the competitiveness of MI w.r.t performance vs cost trade-off.

### 3.2. Results of the Bayesian Framework

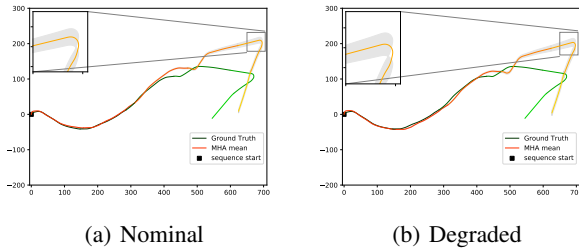


Figure 3. Uncertainty in some poses of trajectory for Sequence 10.

Next, we evaluate uncertainty estimates in the form of variance in mean predicted poses, which is depicted with the grey shaded region around the trajectory in Figure 3. From Figure 3b it can be seen that increased variance expresses the models’ uncertainty about the point. To further evaluate the behavioral effects of uncertainty, some more representative results can be seen in Figure 4. The errors and uncertainties

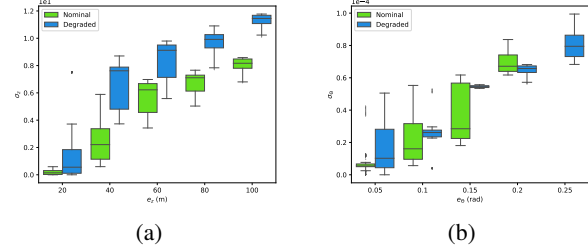


Figure 4. Box plots of uncertainty vs. errors for translation along z direction and pitch angle  $\theta$  respectively for Sequence 10.

are shown for nominal (green) and occlusion degradation (blue) cases. For Sequence 10 it was observed that the pose error was large along Z-axis as compared to others. A large uncertainty can be seen along this axis which suggests that the error fluctuations can be better captured by it. The box plots show steady increase in uncertainty along with error (strong correlation with error). Higher uncertainty is observed for degraded case in all plots. Lastly, end-to-end pose regression approaches (Kendall et al., 2015) such as ours, are often used in the hybrid methods such as (Yang et al., 2020) that effectively leverage the domain knowledge within the learner. In this sense, purely a multi-modal learning approaches can also be useful for such advanced methods beyond end-to-end strategy. For future works, further evaluations on other public datasets (Sun et al., 2019; Caesar et al., 2019) and other applications of VIO (Lutz et al., 2020; Lee et al., 2020a) can be envisioned.

## 4. Conclusion

We introduced a learning based method for VIO, in which, sensor fusion was carried out using multi-head self-attention. Our approach was evaluated for the scenarios with imperfect sensor data, and we demonstrated that robustness improves with a more expressive fusion strategy. Employing Laplace Approximation, we made our model capable of making predictions with a notion of uncertainty, and demonstrated that the errors in the pose and their confidence can be made correlated for end-to-end VIO based approaches.

## Acknowledgements

We thank the anonymous reviewers and the area chairs for their time and thoughtful comments. The authors acknowledge the support by the Helmholtz Association’s Initiative and Networking Fund (INF) under the Helmholtz AI platform grant agreement (contract number ID ZT-I-PF-5-1), the project ARCHES (contract number ZT-0033) and the EU-project AUTOPILOT (contract number 731993).

## References

- Almalioglu, Y., Turan, M., Sari, A. E., Saputra, M. R. U., de Gusmão, P. P., Markham, A., and Trigoni, N. Selfvio: Self-supervised deep monocular visual-inertial odometry and depth estimation. *arXiv preprint arXiv:1911.09968*, 2019.
- Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019.
- Chen, C., Lu, X., Markham, A., and Trigoni, N. Ionet: Learning to cure the curse of drift in inertial odometry. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Chen, C., Rosa, S., Miao, Y., Lu, C. X., Wu, W., Markham, A., and Trigoni, N. Selective sensor fusion for neural visual-inertial odometry. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- Clark, R., Wang, S., Markham, A., Trigoni, N., and Wen, H. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017a.
- Clark, R., Wang, S., Wen, H., Markham, A., and Trigoni, N. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017b.
- Couzinié-Devy, F., Sun, J., Alahari, K., and Ponce, J. Learning to estimate and remove non-uniform image blur. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1075–1082, 2013.
- Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. FlowNet: Learning optical flow with convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- El-Sheimy, N., Hou, H., and Niu, X. Analysis and modeling of inertial sensors using allan variance. *IEEE Transactions on instrumentation and measurement*, 57(1):140–149, 2007.
- Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.
- Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- Geiger, A., Lenz, P., Stiller, C., and Urtasun, R. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
- Han, L., Lin, Y., Du, G., and Lian, S. Deepvio: Self-supervised deep learning of monocular visual inertial odometry using 3d geometric constraints. *arXiv preprint arXiv:1906.11435*, 2019.
- Jayakumar, S. M., Czarnecki, W. M., Menick, J., Schwarz, J., Rae, J., Osindero, S., Teh, Y. W., Harley, T., and Pascanu, R. Multiplicative interactions and where to find them. In *International Conference on Learning Representations*, 2020.
- Kendall, A. and Cipolla, R. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4762–4769, May 2016. doi: 10.1109/ICRA.2016.7487679.
- Kendall, A. and Cipolla, R. Geometric loss functions for camera pose regression with deep learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- Kendall, A., Grimes, M., and Cipolla, R. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.
- Lee, J., Balachandran, R., Sarkisov, Y. S., De Stefano, M., Coelho, A., Shinde, K., Kim, M. J., Triebel, R., and Kondak, K. Visual-inertial telepresence for aerial manipulation. *arXiv preprint arXiv:2003.11509*, 2020a.
- Lee, J., Humt, M., Feng, J., and Triebel, R. Estimating model uncertainty of neural networks in sparse information form. *arXiv preprint arXiv:2006.11631*, 2020b.
- Li, R., Wang, S., Long, Z., and Gu, D. Undeepvo: Monocular visual odometry through unsupervised deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7286–7291, May 2018. doi: 10.1109/ICRA.2018.8461251.

- Lutz, P., Müller, M. G., Maier, M., Stoneman, S., Tomić, T., von Bargaen, I., Schuster, M. J., Steidle, F., Wedler, A., Stürzl, W., et al. Ardea—an mav with skills for future planetary missions. *Journal of Field Robotics*, 37(4): 515–551, 2020.
- MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.
- Ritter, H., Botev, A., and Barber, D. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.
- Shamwell, E. J., Leung, S., and Nothwang, W. D. Vision-aided absolute trajectory estimation using an unsupervised deep network with online error correction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2524–2531, Oct 2018.
- Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.
- Sun, D., Yang, X., Liu, M.-Y., and Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.
- Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., and Anguelov, D. Scalability in perception for autonomous driving: Waymo open dataset, 2019.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.
- Walch, F., Hazirbas, C., Leal-Taixe, L., Sattler, T., Hilsenbeck, S., and Cremers, D. Image-based localization using lstms for structured feature correlation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- Wang, S., Clark, R., Wen, H., and Trigoni, N. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2043–2050, 2017.
- Wang, S., Clark, R., Wen, H., and Trigoni, N. End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks. *The International Journal of Robotics Research*, 37(4-5):513–542, 2018.
- Wang, T.-C., Efros, A. A., and Ramamoorthi, R. Occlusion-aware depth estimation using light-field cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3487–3495, 2015.
- Yang, N., Stumberg, L. v., Wang, R., and Cremers, D. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1281–1292, 2020.
- Yin, Z. and Shi, J. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- Zhan, H., Garg, R., Saroj Weerasekera, C., Li, K., Agarwal, H., and Reid, I. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- Zhou, T., Brown, M., Snavely, N., and Lowe, D. G. Unsupervised learning of depth and ego-motion from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.



# Supplementary Material

## 1. Related Work

We consider Deep learning (DL) based approaches for visual localization. Amongst many, we review the most related work as follows. Wang et al. (2017) describes an end-to-end trainable approach without using modules from conventional VO pipeline. One of the first deep learning approaches that used CNNs for estimation of 6 DoF poses is PoseNet (Kendall et al., 2015). Some works have shown improved performance by making certain modifications to PoseNet, for e.g. in Kendall & Cipolla (2017) by changing the loss function to incorporate geometric information. Addition of LSTM after CNN similar to Wang et al. (2017) is made to the network in (Walch et al., 2017; Clark et al., 2017a) and the network is further improved by making the CNN part Bayesian to estimate uncertainty of localization in Kendall & Cipolla (2016). Probabilistic approach has also been explored in Wang et al. (2018). Several approaches (Li et al., 2018; Zhan et al., 2018; Zhou et al., 2017; Yin & Shi, 2018) have used unsupervised learning approach for depth estimation and VO.

In learning based methods for VIO like VINet (Clark et al., 2017b), wherein the two features streams are simply concatenated and fed into another LSTM for pose regression. It is one of the early end-to-end frameworks for VIO trained in a supervised manner. The work in Chen et al. (2019), replaces the simple fusion strategy used in Clark et al. (2017b) with a more sophisticated one. Two approaches are suggested - soft and hard fusion which do the fusion either deterministically or stochastically. VIOLearner (Shamwell et al., 2018) is an unsupervised scaled trajectory estimation and online error correction work. However, the necessity of depth information to recover absolute scale may make it difficult to use as it may not always be available. Self-supervised approaches like DeepVIO (Han et al., 2019) use optical flow and pre-integrated IMU network with a status update module to update its status, and SelfVIO (Almalioglu et al., 2019) does monocular VIO and depth reconstruction using adversarial training.

## 2. Implementation details

### 2.1. Dataset

The KITTI odometry benchmark (Geiger et al., 2012; 2013) for odometry evaluations consists of 22 image sequences.

Images are acquired at the rate of 10 Hz and saved in png format. OXTS RT 3003 is a GPS/IMU localization system that captures IMU data at 100 Hz (found in raw KITTI dataset) and also determines the ground truth data for this dataset. The ground truth is also acquired at 10 Hz. Sequences 00-10 have corresponding ground truth data associated with them for training whereas sequences 11-22 are without corresponding ground truth for evaluation purposes. We use the sequences 00, 01, 02, 05, 08, 09 for training because they are comparatively longer and sequences 04, 06, 07, 10 for testing. Sequence 03 is omitted due to missing raw data file. Images and raw IMU data are manually synchronized. In order to generate more training data, the sequences are segmented into trajectories of varying lengths which in turn help in avoiding overfitting by core LSTM. Here the varying sequence length is treated as a hyperparameter.

### 2.2. Architecture

The network is implemented using PyTorch framework (Paszke et al., 2019) and trained on an NVIDIA GTX 1080 GPU. The Adam optimizer with learning rate of  $5e^{-5}$  is used to train the network for up to 80 epochs with a batch size of 8. The scale factor  $\beta$  for MSE is chosen to be 1000 as improved orientation estimates are obtained at this value. Pre-trained FlowNetS model (Dosovitskiy et al., 2015) is used for the CNN and transfer learning is done from our implementation of the VINet model for both feature extractors to reduce time required for training and convergence. Regularization techniques like batch-normalization and dropout have been used. For the FlowNet adaptation, the CNN has total 17 layers, each convolutional layer followed by ReLU activation except for last convolutional layer. Images have been resized to  $184 \times 608$ . The IMU-LSTM and the core LSTM both are chosen to be bidirectional with two layers with 15 and 1000 hidden states respectively.

### 2.3. Baselines

For all the baselines, as open source code was not available, and thus, the implementations might differ. The implementation details are as follows. The same configuration for FlowNet and LSTM has been followed for the baselines since a performance improvement was found. specific choice stems For DeepVO, Adagrad optimizer with a learning rate of  $5e^{-4}$  is used, the network is trained for 250

epochs. VINet is trained using Adam optimizer with learning rate of  $1e^{-4}$ . We have removed the SE(3) composition layer used in original work and replaced it with a fully-connected layer. VIO Soft is trained using Adam optimizer with a learning rate of  $1e^{-5}$ .

### 3. Degradation datasets

The sensor corruptions tell a lot about network robustness. Robustness of the network to such degradations provides an insight into the underlying sensor fusion mechanism. It points towards the resilience of the network to sensor corruption and consequent failure, highlighting the complementary nature of the sensor modalities in face of adversity. It also gives an insight into how different fusion strategies give relative importance to visual and inertial features.

For this reason, we prepare two groups of datasets by introducing various types of sensor degradation as done in (Chen et al., 2019). They are as under:

#### 3.1. Visual-Inertial degradation

- Part occlusion: We cutout a part of dimension  $200 \times 200$  pixels from an image at a random location by overlaying a mask of the same size on it. Such situation can occur when camera view is obstructed by objects very close to it or due to dust (Wang et al., 2015).
- Noise: We add salt and pepper noise along with Gaussian blur to the images. This can occur due to substantial horizontal motion of the camera, changing light conditions or due to defocusing (Couzinié-Devry et al., 2013).
- Missing frames: Some images are removed at random. This happens when the sensor temporarily gets disconnected or while passing through low-lit area like a tunnel.
- Noise: IMU has inherent errors due to biases in accelerometers and gyroscopes and gyro drift. In addition to the already existing noise we add white noise to accelerometer and bias to gyroscope. This can happen due to temperature variations giving rise to white noise and random walk (El-Sheimy et al., 2007).
- Missing frames: Random removal of IMU frames between two image frames is done. This is plausible due to packet loss from bus or sensor instability.

### 4. Additional results for sensor degradation

Tables 2, 3, 4 and 5 are the detailed version of Table 1 comparing the metric for sequences 04, 06, 09 and 10. Figure 5 and Figure 6 substantiate the analysis presented in Section 3.1.

Table 2. Comparison metric for nominal case

SEQ	DEEPVO		VINET		VIO SOFT		MHA	
	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$
04	<b>6.87</b>	2.48	11.30	2.63	10.84	2.42	8.90	<b>1.58</b>
06	26.52	8.90	<b>15.33</b>	<b>5.25</b>	17.75	6.91	15.59	5.56
09	18.63	4.68	<b>6.95</b>	<b>1.79</b>	6.98	2.26	8.38	2.80
10	23.1	<b>4.33</b>	22.34	8.98	19.03	7.74	<b>12.82</b>	5.80
MEAN	18.07	5.88	13.98	4.66	13.65	4.83	<b>11.42</b>	<b>3.94</b>

Table 3. Comparison metric for IMU degradation case

SEQ	DEEPVO		VINET		VIO SOFT		MHA	
	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$
04	N/A	N/A	12.56	3.02	10.563	2.95	<b>9.30</b>	<b>1.60</b>
06	N/A	N/A	<b>16.71</b>	<b>5.31</b>	18.87	7.16	17.04	5.82
09	N/A	N/A	10.47	3.50	7.71	2.33	<b>7.40</b>	<b>2.13</b>
10	N/A	N/A	25.48	9.28	19.07	7.33	<b>16.18</b>	<b>6.15</b>
MEAN	-	-	16.31	5.28	14.05	4.94	<b>12.48</b>	<b>3.93</b>

Table 4. Comparison metric for all vision degradation case

SEQ	DEEPVO		VINET		VIO SOFT		MHA	
	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$
04	F	F	12.29	2.83	11.82	2.43	<b>9.61</b>	<b>1.90</b>
06	A	A	<b>14.80</b>	<b>5.08</b>	19.01	6.78	16.75	5.67
09	I	I	10.11	3.39	<b>7.44</b>	<b>2.20</b>	10.22	2.84
10	L	L	24.67	8.55	20.06	6.92	<b>13.86</b>	<b>5.53</b>
MEAN	-	-	15.46	4.97	14.59	4.58	<b>12.61</b>	<b>3.98</b>

Table 5. Comparison metric for all sensor degradation case

SEQ	DEEPVO		VINET		VIO SOFT		MHA	
	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$	$t_{rel}$	$r_{rel}$
04	N/A	N/A	13.59	3.09	12.98	2.61	<b>10.67</b>	<b>2.03</b>
06	N/A	N/A	<b>15.00</b>	<b>4.98</b>	16.94	5.91	15.38	5.17
09	N/A	N/A	10.39	3.50	<b>9.34</b>	2.71	9.93	<b>2.66</b>
10	N/A	N/A	24.49	9.00	22.65	7.28	<b>13.88</b>	<b>5.30</b>
MEAN	-	-	15.87	5.14	15.48	4.63	<b>12.47</b>	<b>3.79</b>

### 5. Laplace Approximation

Given model parameters  $\theta$  and data  $\mathcal{D}$ , the log posterior can be captured approximately by a second order Taylor expansion around the maximum a posteriori (MAP) estimate,

$$\log p(\theta|\mathcal{D}) \approx \log p(\theta_{MAP}|\mathcal{D}) - \frac{1}{2}(\theta - \theta_{MAP})^T H(\theta - \theta_{MAP}),$$



where  $H$  is the matrix of second order derivatives (Hessian) w.r.t. the negative log posterior. Notably, the right hand side is of Gaussian functional form when applying the exponential, thus it can be used to sample parameter configurations:

$$\theta \sim \mathcal{N}(\theta_{MAP}, H^{-1}).$$

Because the Hessian of a deep neural network is infeasible to compute and invert, approximations to the Fisher information matrix (Fisher), like Kronecker factored or diagonal approximations, are used instead.

Approximate Bayesian inference can be performed using Monte Carlo integration, by sampling multiple parameter configurations from the approximate posterior and collecting the model output for each of them given an unseen datum  $\tilde{\mathbf{x}}$ :

$$\begin{aligned} p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \mathcal{D}) &= \int p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta) p(\theta|\mathcal{D}) d\theta \\ &\approx \frac{1}{T} \sum_{t=1}^T p(\tilde{\mathbf{y}}|\tilde{\mathbf{x}}, \theta_t). \end{aligned}$$

The mean is used to make predictions while for regression, the variance is a measure of the predictive uncertainty. We make use of  $T = 30$  posterior samples in all our experiments. Because of numerical instability, overestimation of variance and wrong attribution of probability mass, the posterior approximation needs to be regularized in practice. We make use of the following simple regularization scheme:

$$NF + \tau I,$$

where  $F$  is the approximation to the Fisher and  $N$  and  $\tau$  are treated as hyperparameters to be optimized on a validation set. We made use of our own implementation of LA for Bayesian inference in PyTorch which we will make publicly available in <https://github.com/DLR-RM>.

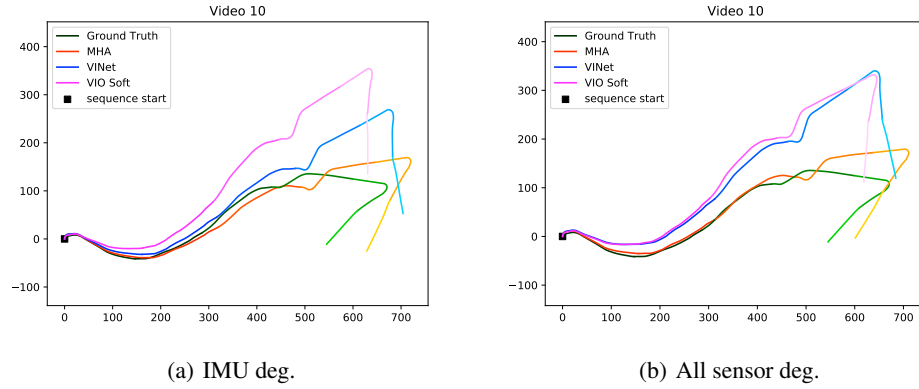


Figure 5. Predicted trajectories for (a) IMU degradation (b) all sensor degradation for Sequence 10.

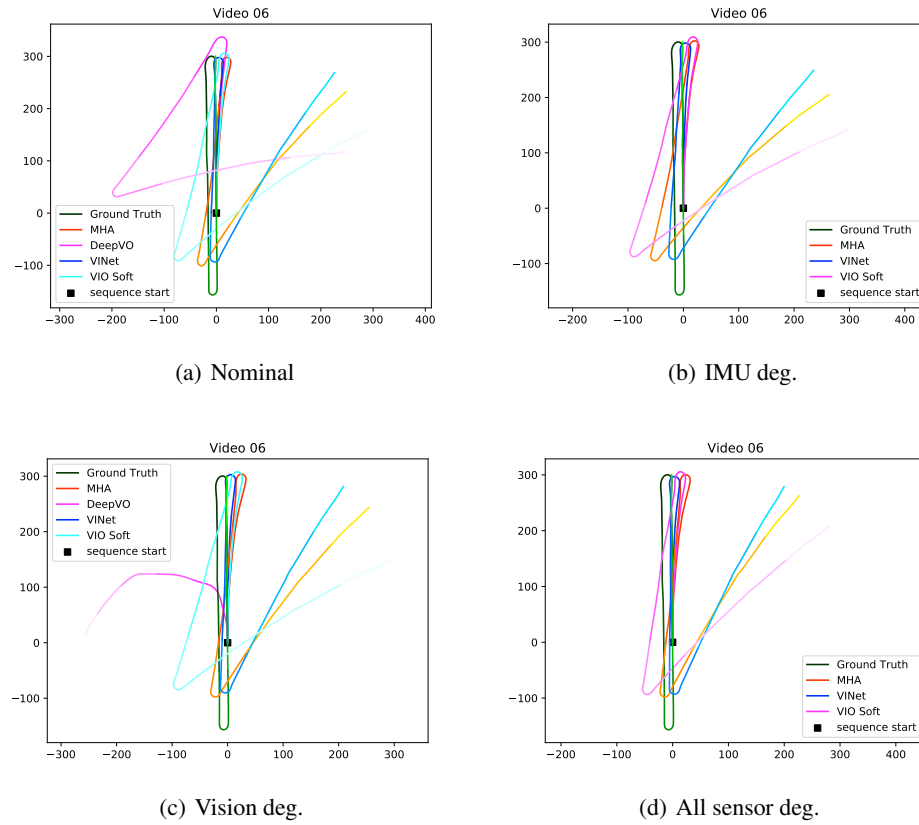


Figure 6. Predicted trajectories with corresponding datasets for Sequence 06.